AD 703924

CNA

ESTIMATION OF BINOMIAL PARAMETERS FROM SEARCH DATA

By

J. Bram and H. Weingarten

Research Contribution No. 3

CONTRACT NONR 3732(00)

Research Contribution

OPERATIONS RESEARCH & MATHEMATICAL SCIENCES DIVISION

Center for Naval Analyses

WASHINGTON 25, D. C.

10

5 May 1964

RESEARCH CONTRIBUTION

Operations Research & Mathematical Sciences Division

CENTER FOR NAVAL ANALYSES

ESTIMATION OF BINOMIAL PARAMETERS FROM SEARCH DATA

By

J. Bram and H. Weingarten

Research Contribution No. 3

J. Bram

H. Weingarten

ABSTRACT

The parameters of the binomial distribution are the constant probability from trial to trial, and the number of trials. In sampling from a binominal population, the number of trials is usually known. Here ~~we provide~~ estimates, are given from observed data, of the constant probability and (because it is also unknown) of the number of trials.

I. INTRODUCTION

The standard binomial sampling problem involves repeating an experiment m times and counting the number of successes. If the random variable representing the number of successes is X, then when we assume X has a binomial distribution we are saying

$$P\left\{X\right\} = P\left\{X = x\right\} = \frac{m!}{x!\,(m - x)!}\, p^x q^{m - x}$$

where $q = 1 - p$ and x may take the values 0, 1, 2,...,m. Here $P\left\{X\right\}$ stands for the distribution function of X or the probability that X takes a particular value. From the observed number of successes in m trials, we are interested in obtaining an estimate of the unknown probability, p, assumed to be constant from trial to trial. The well known answer to this estimation problem is to estimate p by

$$\hat{p} = \frac{s}{m}$$

where s is the number of successes in m trials. This estimate has been shown to have the desirable properties of unbiasedness, and minimum variance. The binomial sampling problem discussed in this paper is different in that in addition to estimating p, we also need an estimate of the number of trials.

II. THE PROBLEM

During a search we assume that there are a fixed but unknown number of objects, m, available to be detected. We assume that the probability of detecting an object is an unknown constant p. We repeat the search k times. All objects detected in the first search are tagged in some manner so that during the second search all new detections will be recognized and tagged. This is repeated for all k searches. We record after each search only the new detections made during that search. We then have k random variables

$$d_j,\ j = 1,\ 2, \ldots,\ k$$

where d_j is the number of new detections made during the j-th search. We want to estimate p and m from the observed values of d_j.

III. RESULTS

The search problem given above describes sampling from a binomial population. We easily see that the distribution of d_1 is given by

$$P(d_1) = \frac{m!}{d_1!\,(m - d_1)!}\, p^{d_1} q^{m - d_1}.$$

The distribution of d_2 given the value of d_1 is then seen to be

$$P(d_2 \mid d_1) = \frac{(m - d_1)!}{d_2!\,(m - d_1 - d_2)!} p^{d_2} q^{m - d_1 - d_2} .$$

In general, then, the distribution of d_j given the $d_1, d_2, \ldots d_{j-1}$ is

$$P(d_j \mid d_1, d_2, \ldots, d_{j-1}) = \frac{(m - d_1 - d_2 - \ldots - d_{j-1})!}{d_j!\,(m - d_1 - d_2 - \ldots - d_j)!} p^{d_j} q^{m - d_1 - d_2 - \ldots - d_j} . \tag{1}$$

From the above, it follows that the joint density of all the d_j, $j = 1, \ldots, k$ the product of all the conditional densities given above, is

$$P(d_1, d_2, \ldots, d_k) = \frac{m!}{d_1!\,d_2! \ldots d_k!\,(m - d_1 - d_2 \ldots d_k)!} p^S q^{k(m - S) + T - S} \tag{2}$$

where

$$S = \sum_{j=1}^{k} d_j$$

$$T = \sum_{j=1}^{k} j d_j .$$

The joint characteristic function of the d_j is the expected value

$$E\left\{ e^{i\left[d_1 t_1 + d_2 t_2 + \ldots + d_k t_k\right]} \right\}$$

and is

$$\sum_{d_1}\sum_{d_2}\cdots\sum_{d_k} \frac{m!}{d_1!d_2!\ldots d_k!(m - d_1 - d_2 - \ldots - d_k)!}\left(pe^{it_1}\right)^{d_1}$$

$$\times \quad \left(pe^{it_2}\right)^{d_2}\ldots\left(pe^{it_k}\right)^{d_k} q^{k(m - S) + T - S} \quad .$$

Performing the summations gives

$$\varphi(t_1, t_2, \ldots, t_k) = \left(pe^{it_1} + pqe^{it_2} + \ldots + pq^{k-1}e^{it_k} + q^k\right)^m \quad . \tag{3}$$

Now the characteristic function of a binomial random variable defined by the parameters N, P (N trials, P constant probability from trial to trial)

is

$$\varphi(T) = \left[Pe^{iT} + Q\right]^N \quad . \tag{4}$$

If in (3) we set all the t's equal to zero except t_j, we see that

$$\varphi(0, 0, \ldots, t_j, 0, 0, \ldots, 0) = \left(p + pq + \ldots + pq^{j-2} + pq^{j-1}e^{it_j} + \right.$$

$$\left. pq^j + \ldots + pq^{k-1} + q^k\right)^m = \left(pq^{j-1}e^{it_j} + (1 - pq^{j-1})\right)^m \quad .$$

By analogy with (4) we then have that the distribution of the individual d_j is

$$P(d_j) = \frac{m!}{d_j!(m - d_j)!}\left(pq^{j-1}\right)^{d_j}\left(1 - pq^{j-1}\right)^{m - d_j} \quad . \tag{5}$$

By setting in (3) all the t_j equal to t we find

$$(t, t, \ldots, t) = \left(\left[p + pq + \ldots + pq^{k-1}\right]e^{it} + q^k\right)^m$$

$$= \left(\left[1 - q^k\right]e^{it} + q^k\right)^m ,$$

and so again by analogy with (4) we see that the distribution of $S = \Sigma d_j$ is given by

$$P_m(S) = \frac{m!}{S!\,(m-S)!}\left(1 - q^k\right)^S \left(q^k\right)^{m-S} . \tag{6}$$

From (2) we may find the maximum likelihood estimate of p by differentiating log $P(d_1, d_2, \dots, d_k)$ with respect to p and finding the value of p which makes this derivative equal to zero. This leads to

$$\hat{p} = \frac{S}{k(m-S)+T} . \tag{7}$$

Again starting with (2) it is possible to show that S is sufficient for m. This follows from the fact that (2) may be rewritten as

$$P(d_1, \dots, d_k) = P(S,\, m)\, P(x_1, x_2, \dots, x_{k-1}) \tag{8}$$

$$= \frac{m!}{S!\,(m-S)!}\left(1 - q^k\right)^S \left(q^k\right)^{m-S} \cdot P(x_1, \dots, x_{k-1})$$

where

$$P(x_1, \dots x_{k-1}) = \frac{S!}{x_1! \dots x_{k-1}!\,(S - x_1, - \dots x_{k-1})!} \frac{p^S}{(1-q^k)^S}$$

$$X_q \quad (k-1)\,S + \sum_{j=1}^{k-1} (j-k)x_j .$$

and $P\,(S,\, m)$ is the same as $P_m(S)$ defined in (6).

The first factor on the right side of (8) is a density as was already seen in (6). The second factor is also a density as seen from

$$\sum P(x_1, \ldots, x_{k-1}) = \left(\frac{p}{1-q^k}\right)^S$$

$$\times \sum_{x_1} \sum_{x_2} \sum_{x_{k-1}} \frac{S!}{x_1! \ldots x_{k-1}! \, (S - x_1 - \ldots - x_{k-1})!} \, q^{(k-1)S + \sum_1^{k-1} (j-k) x_j}$$

$$= \left(\frac{p}{1-q^k}\right)^S \left(\frac{1-q^k}{1-q}\right)^S = 1 \, ,$$

in which the sum on the left is taken over-all $(x_1, \ldots, x_{k-1})$ such that $\sum_{j=1}^{k-1} x_j = S$.

The second factor on the right in (8) comes from the transformation

$$\begin{aligned} x_1 &= d_1 \\ x_2 &= d_2 \\ &\vdots \\ x_{k-1} &= d_{k-1} \\ S &= d_1 + d_2 + \ldots + d_k \end{aligned}$$

applied to (2). The factorization in (8) means that S is sufficient for m. Now we would like to have a function of S as an estimate of m which is unbiased and unique. We want to use a result which says that if a function of a sufficient statistic is both unbiased and unique, it is the best statistic for estimation in the sense of having minimum variance. We propose as this function of S the quantity

$$f(S) = \frac{S}{P} \tag{9}$$

where $P = 1 - q^k$.

Now we want $f(S)$ to be unbiased for all m. This means

$$E\left[f(S)\right] = m, \text{ for all } m \tag{10}$$

or

$$\sum_{S=0}^{m} f(S) P_m(S) = m \text{ for all } m. \tag{11}$$

where $P(S)$ is defined in (6).

We want to prove that $f(S)$ given in (9) is unique. Suppose $S = 0$. We take $m = 0$ and in (11) we have

$$\sum_{S=0}^{0} f(0)P_0(S) = 0$$

or

$$f(0) = 0$$

Suppose $S = 1$. We take $m = 1$ and in (11) we have

$$\sum_{S=0}^{1} f(S)P_1(S) = 1.$$

Now $f(0)$ is already 0 and we find

$$f(1)P_1(1) = 1$$

$$f(1)\frac{1!}{1!0!}P = 1$$

$$f(1) = \frac{1}{P}.$$

Suppose $S = 2$. We take $m = 2$ and require

$$\sum_{S=0}^{2} f(S)P_2(S) = 2$$

We already have $f(0) = 0$, $f(1) = \frac{1}{P}$. Hence we find

$$0 \cdot 1 + \frac{1}{P} \cdot \frac{2!}{1!1!}PQ + f(2) \cdot P^2 = 2$$

or

$$f(2) = \frac{2}{P}.$$

By induction it follows that

$$f(S) = \frac{S}{p},$$

for suppose (9) is correct for $S \le r$. We then have from (11) for $S = r + 1$ and taking $m = r + 1$

$$\sum_{i=0}^{r} \frac{i}{p} P_{r+1}(i) + f(r + 1) P_{r+1}(r+1) = r + 1 \text{ or}$$

$$(r + 1)(1 - p^r) + f(r + 1) p^{r+1} = r + 1 \text{ or}$$

$$f(r + 1) p^{r+1} = (r + 1) p^r$$

and so

$$f(r + 1) = \frac{r+1}{p} .$$

Hence the estimate given in (9) is unique; since it is sufficient and unbiased it has minimum variance. We rewrite (9) as

$$\hat{m} = \frac{S}{1-q^k} . \tag{12}$$

This estimate of m (12), together with

$$\hat{p} = \frac{S}{k(\hat{m}-S)+T} \tag{7}$$

may be solved simultaneously for $\hat{m}$ and $\hat{p}$.